ABSTRACT OF THE DISCLOSURE

A method and apparatus for extracting information from symbolically compressed document images. A deciphering module generates first and second text strings by deciphering respective sequences of template identifiers in first and second symbolically compressed document images. A conditional n-gram module receives the first and second text strings from the deciphering module and extracts n-gram terms therefrom based on a predicate condition. A comparison module generates a measure of similarity between the first and second symbolically compressed document images based on the n-gram terms extracted by the conditional n-gram module.